Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 105-108 recibido 14-03-2020 revisado 30-05-2020 aceptado 30-05-2020

AMALEU: Una Representación Universal del Lenguaje basada en Aprendizaje Automático

AMALEU: A Machine-Learned Universal Language Representation

Marta R. Costa-jussà
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
marta.ruiz@upc.edu

Resumen: El objetivo del proyecto AMALEU es aprender una representación común para diferentes idiomas. Se pretende tener una representación común para la lengua oral y una para la lengua escrita. AMALEU, de dos años de duración, está financiado por el MINECO dentro del programa de *Europa Excelencia*
Palabras clave: Traducción Automática Multilingüe, Traducción de Voz y Texto

Abstract: The objective of AMALEU's project is learning a multilingual common representation for speech and a different multilingual common representation for text. AMALEU is a two-year project funded by the MINECO.
Keywords: Multilingual Machine Translation, Speech and Text Translation

1 Participantes del proyecto

El grupo de investigación que participa en el proyecto pertenece al grupo de investigación de Voz del Departamento de Teoría de Señal y Comunicaciones de la Universidad Politècnica de Cataluña y al centro de investigación TALP. La investigadora principal es la autora de este artículo, y como investigadores están Carlos Escolano, Gerard Gallego y Javier Ferrando.

2 Entidad financiadora

El proyecto está financiado por el Ministerio de Economía y Competitividad y el código del proyecto es EUR2019-103819. AMALEU comenzó el 1 de enero de 2019, finaliza el 31 de enero de 2020 por lo que tiene una duración de 24 meses. La financiación total es de 74.850 euros.

3 Contexto y motivación

¿Por qué la traducción automática entre inglés y portugués es considerablemente mejor que la traducción automática entre holandés y castellano? ¿Por qué los reconocedores automáticos funcionan mejor en alemán que en finés? El principal motivo es la variación en la cantidad de datos etiquetados para entrenar y modelar los sistemas. Aunque el mundo es multimodal y altamente multilingüe, la tecnología de voz y lenguaje no es capaz de absorver la alta demanda que hay en todos los lenguajes. Necesitamos mejores algoritmos de aprendizaje que puedan explotar el progreso de unas pocas modalidades y lenguajes para el beneficio de otros.

Este proyecto se focaliza en el reto de aprendizaje a partir de pocos recursos a partir de una aproximación para la traducción automática multilingüe.

AMALEU propone entrenar conjuntamente un modelo multilingüe y multimodal que aprenda una representación universal del lenguaje. Este modelo compensará la falta de datos etiquetados y mejorará significativamente la capacidad de generalización de los datos de entrenamiento a partir de observar una variedad de recursos no etiquetados. Este modelo reducirá el número cuadrático de sistemas de traducción a linear, lo cual tendrá un gran impacto en un entorno multilingüe.

El reto de este proyecto radica en entrenar una representación universal de manera automática y con aprendizaje profundo. Para esto, AMALEU utilizará una arquitectura basada en la estructura de codificador-decodificador. El codificador aprende una representación del lenguaje de entrada mediante una reducción de dimensionalidad, que será la representación universal del lenguaje; a partir de esta abstracción, el decodificador genera la salida. La arquitectura interna del codificador-decodificador se diseñará explicitamente para aprender la abstracción

ISSN 1135-5948. DOI 10.26342/2020-65-14

universal del lenguaje, que se integrará como función objetivo de la arquitectura.

AMALEU impactará comunidades altamente multidisciplinares incluyendo ciencias de la computación, matemáticas, ingeniería y linguística, que trabajan conjuntamente en aplicaciones del procesado del lenguaje natural de voz y de texto.

4 Proyecto AMALEU

El objetivo de AMALEU es aprender de manera automática una representación universal del lenguaje ya sea voz o texto, de manera que se pueda explotar en aplicaciones de inteligencia artificial. Adicionalmente, el proyecto se plantea utilizar fuentes de información no etiquetadas así como información lingüística. La Figura 1 muestra el diagrama de Gantt del proyecto y los principales paquetes de trabajo. El plan de trabajo incluye: gestión y diseminación, representación multilingüe común, integración de recursos lingüísticos y la integración y evaluación. A continuación, describimos brevemente cada uno de estos cuatro puntos del plan de trabajo, así como el grado de consecución de los objetivos a fecha 1 de junio de 2020 (mes 17).

4.1 Gestión y Diseminación

Esta parte del trabajo incluye la adecuada gestión del proyecto preparando los informes adecuados y controlando el presupuesto adecuadamente. Además, se gestionará una diseminación elaborada que incluya publicaciones y participaciones en eventos internacionales. Las publicaciones relacionadas con el proyecto se van citando cuando se describen la consecución de los objetivos de cada tarea.

Consecución de los objetivos (mes 17). Este objetivo está completado al 75 %, resultados del mismo se pueden ver en acciones de diseminación como el desarrollo de la página web del grupo de investigación y del proyecto[1], la participación en la organización de evaluaciones internacionales (Barrault et al., 2019), y artículos de diseminación del área (Costa-jussà et al., 2020).

4.2 Representación Común de la Voz o el Texto Multilingües

El objetivo en este punto es obtener una representación común del lenguaje en su representación textual y en su representación oral, de manera independiente. La representación común del texto o la voz se construirá a partir de una arquitectura de codificador-decodificador. Dado que múltiples lenguajes se pueden entrenar en un único modelo que pueda producir traducciones multilingües, queremos aprender la representación multilingüe del lenguaje a partir del uso de una función objetivo común.

Dentro de esta parte del trabajo planteamos dos actividades. La primera es el desarrollo de la arquitectura para el lenguaje textual. Aquí nos focalizaremos en investigar mecanismos de atención y en diferentes representaciones intermedias que sean de longitud fija o variable. Evaluaremos tanto la calidad de la representación intermedia como la calidad de la traducción. La segunda actividad consistirá en construir una representación común para el lenguaje oral. Aquí la investigación se centrará en adaptar el codificador de manera que pueda aceptar la entrada oral que es considerablemente larga. Asimismo, identificaremos mecanismos de atención adecuados para tales longitudes. La parte del decodificador se compartirá en ambas actividades.

Consecución de los objetivos (mes 17). La representación común en texto se da por completada a partir de los siguientes trabajos: (Escolano, Costa-jussà, y Fonollosa, 2019; Escolano et al., 2019; Escolano et al., 2020a; Escolano et al., 2020b), donde se demuestra que la arquitectura que hemos planteado es capaz de acercar las representaciones de oraciones similares en distintos idiomas. La parte de voz está en curso.

4.3 Recursos Lingüísticos

El objetivo en este punto es utilizar información lingüística externa de manera que ayude a encontrar la representación común del lenguaje. Para eso se investigará sobre cuales son las mínimas unidades (palabras, subpalabras o caracteres) más adecuadas para conseguir una representación universal del lenguaje. Por otro lado se explotarán recursos de diferentes naturalezas incluyendo bases de datos monolingües.

Consecución de los objetivos (mes 17). Hemos explorado las unidades mínimas y información lingüística (Casas, Costa-jussà, y Fonollosa, 2020; Casas, Fonollosa, y Costa-jussà, 2020; Armengol-Estapé, Costa-jussà, y Escolano, 2020) así como el uso de datos monolingües (Biesialska, Rafieian, y Costa-

[1] http://mt.cs.upc.edu

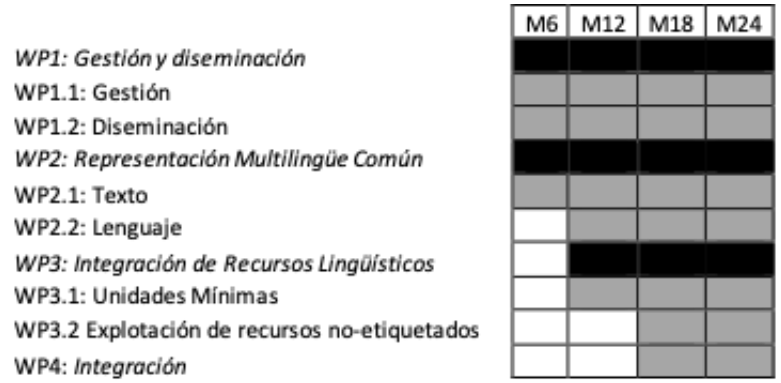

Figura 1: Diagrama Gantt. Plan de trabajo

jussà, 2020). Ahora falta su integración en la arquitectura multilingüe.

4.4 Integración/Evaluación

En cuando a datos, AMALEU usará datos multimodales de bases de datos existentes y conocidas como los que provienen de las evaluaciones del WMT[2] y del IWSLT[3] así como otros recursos disponibles desde el Linguistic Data Consortium[4].

Referente a la evaluación, como no existe un método establecido para evaluar la calidad de la representación intermedia, proponemos una nueva medida que utiliza la distancia de la representación entre oraciones similares. Esta distancia se combina con la evaluación de recuperación de la oración original a partir de su representación para calcular lo que denominamos Medida de Similitud con Recuperación. Asimismo, calcularemos con oraciones que son contradictorias, tienen una representación no similar (a partir de la medida de distancia entre oraciones) y combinaremos esta distancia con la recuperación de la frase original para calcular lo que denominamos Medida de Disimilitud con Recuperación. Estas medidas pretenden controlar que las oraciones similares tengan una representación similar y las oraciones contradictorias tengan una representación distante. La calidad de la traducción se evaluará con el método estandard BLEU (Papineni et al., 2002).

Consecución de los objetivos (mes 17). La transversalidad de esta tarea hace que su consecución se mida mayoritariamente en base a los objetivos previos. Además, queremos añadir que con nuestro sistema hemos participado en evaluaciones internacionales de prestigio (Casas et al., 2019) obteniendo siempre resultados competitivos.

5 Impacto

El impacto de AMALEU se refleja en una mejora de calidad y eficiencia de los actuales sistemas multilingües más allà de los traductores automáticos.

La novedosa integración de información multilingüe y multimodal en las aplicaciones de inteligencia artificial, que plantea AMALEU, permitirá mejorar la calidad de las mismas. Es clave que los sistemas pueden explotar múltiples recursos y se puedan entrenar a partir de datos no etiquetados. Como consecuencia, se consigue un aprendizaje zero-shot, que quiere decir que el sistema no necesita ver ejemplos de la tarea en concreto para aprenderla.

La extensión del entrenamiento a múltiples lenguas y modalidades contribuirá a mayores generalizaciones. Asimismo, la explotación de una alta variedad de recursos lingüísticos también lo hará.

La representación universal del lenguaje permite abrir nuevos horizontes en tareas como la búsqueda de información o los resúmenes automáticos cross-lingües. Asimismo, el uso de una representación común del lenguaje permitirá reducir el número de sistemas multilingües de N·(N-1) (donde N es el número de lenguas) a 2·N. Además nuestra aproximación permite añadir incrementalmente nuevas lenguas. Este logro consigue crear nuevas arquitecturas de aprendizaje profundo más efi-

[2]http://www.statmt.org/wmt20/
[3]https://workshop2020.iwslt.org/
[4]https://www.ldc.upenn.edu/

cientes, así como beneficiar las lenguas de pocos recursos a partir de las lenguas con más recursos.

Finalmente, mencionar que la investigación realizada en AMALEU tiene en cuenta producir resultados que sean justos y no reproduzcan sesgos sociales (Costa-jussà, 2019; Costa-jussà, Lin, y España-Bonet, 2020; Costa-jussà et al., 2020).

Agradecimientos

Este trabajo esta financiado por el Ministerio de Economía y Competitividad a través del projecto EUR2019-103819 y el programa Ramón y Cajal que incluye financiamiento del European Regional Development Fund.

Bibliografía

Armengol-Estapé, J., M. R. Costa-jussà, y C. Escolano. 2020. Enriching the transformer with linguistic and semantic factors for low-resource machine translation. *ArXiv*, abs/2004.08053.

Barrault, L., O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post, y M. Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). En *Proceedings of the WMT*, páginas 1–61, Florence.

Biesialska, M., B. Rafieian, y M. R. Costa-jussà. 2020. Enhancing word embeddings with knowledge extracted from lexical resources. En *ACL Student Research Workshop*.

Casas, N., M. R. Costa-jussà, y J. A. R. Fonollosa. 2020. Combining subword representations into word-level representations in the transformer architecture. En *ACL Student Research Workshop*.

Casas, N., J. A. R. Fonollosa, y M. R. Costa-jussà. 2020. Syntax-driven iterative expansion language models for controllable text generation. *ArXiv*, abs/2004.02211.

Casas, N., J. A. R. Fonollosa, C. Escolano, C. Basta, y M. R. Costa-jussà. 2019. The TALP-UPC machine translation systems for WMT19 news translation task: Pivoting techniques for low resource MT. En *Proceedings of the WMT*, Florence, Italy.

Costa-jussà, M. R., R. Creus, O. S. Domingo, A. S. Dominguez, M. Escobar, C. I. López, M. Y. F. García, y M. Geleta. 2020. Mt-adapted datasheets for datasets: Template and repository. *ArXiv*, abs/2005.13156.

Costa-jussà, M. R. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1(11):495–496.

Costa-jussà, M. R., C. España-Bonet, P. Fung, y N. A. Smith. 2020. Multilingual and interlingual semantic representations for natural language processing: A brief introduction. *Computational Linguistics*.

Costa-jussà, M. R., P. L. Lin, y C. España-Bonet. 2020. Gebiotoolkit: Automatic extraction of gender-balanced multilingual corpus of wikipedia biographies. En *Proc of the LREC*.

Escolano, C., M. R. Costa-jussà, y J. A. R. Fonollosa. 2019. From bilingual to multilingual neural machine translation by incremental training. En *Proceedings of the ACL Student Research Workshop*, páginas 236–242, Florence.

Escolano, C., M. R. Costa-jussà, J. A. R. Fonollosa, y M. Artetxe. 2020a. Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. *ArXiv*, abs/2004.06575.

Escolano, C., M. R. Costa-jussà, J. A. R. Fonollosa, y M. Artetxe. 2020b. Training multilingual machine translation by alternately freezing language-specific encoders-decoders. *ArXiv*, abs/2006.01594.

Escolano, C., M. R. Costa-jussà, E. Lacroux, y P.-P. Vázquez. 2019. Multilingual, multi-scale and multi-layer visualization of intermediate representations. En *Proceedings of the EMNLP-IJCNLP: System Demonstrations*, Noviembre.

Papineni, K., S. Roukos, T. Ward, y W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. En *Proceedings of ACL*, páginas 311–318, Philadelphia, Pennsylvania, USA, Julio.